



YEARS 12/13

HL IB

VISIT...

IB

M

LANZAROTE
Caliente.COM

ATICS

Op

ty

2nd Edition

IB MATH HIGHER OPTION

STATISTICS

There is no doubt that the options in the Higher syllabus are a stern test of both your knowledge and understanding of more advanced mathematics than you will meet in the core syllabus. This revision booklet has been designed to cover most aspects of the option syllabus along with a variety of examples which show how the concepts can be applied. I cannot, however, hope to cover every aspect of every topic in this booklet, and it is to be regarded as a supplement to such text books and notes as you have been given during the course. Do note that calculator references in this book are to the TI-84 family. They are still useful even if you have a different calculator as they indicate what is possible. Your textbook or an online instruction manual should explain how you can do similar operations on your calculator.

Many candidates fall down on the option questions because they have not put enough preparation in. The practice questions after each section test your knowledge of the syllabus and you should try to answer all of them. If you have difficulty with any of them, talk to your teacher or fellow students, read your notes and textbooks, and then come back and try again.

Your teacher will no doubt give you plenty of questions taken from past examination papers. Do be aware as you go through these that any papers sat before 2014 will not include questions on the efficiency of estimators, probability generating functions, correlation or linear regression.

The following ideas should help you answer questions with more accuracy and more confidence:

- The option syllabus progresses logically from the probability and statistics in the core. It is important that you fully understand these basic principles, otherwise you will never fully grasp the more advanced concepts in this option.
- There is a great deal of numerical work required, particularly when answering statistics questions. You are doing yourself a disservice if you get the method right, but cannot compute the correct answers. Get into the habit of double-checking every calculation; and make sure you show all your working clearly so that if the answer is wrong, you can pick up method marks.
- Your calculator is your friend, so you should be aware of all the ways in which it can help you. Make sure you practise answering questions using your calculator – but you *must* understand the underlying mathematics as well.

Discrete Distributions

A probability distribution shows the probabilities for all the outcomes of a particular event. Discrete probability distributions relate to events which can only have certain outcomes – usually with integer values. Some discrete distributions are covered in the core syllabus: the **uniform** distribution, the **Binomial** distribution, the **Poisson** distribution and those defined by a **formula** – often called *probability mass functions*. You will find these described in the HL revision book. Then there are:

In each of these distributions, $P(\text{success})$ is written as p .

Geometric distribution: The number of attempts until the first success (where, each trial outcome is either success or failure). The probability of the first success on the n th attempt is $(1 - p)^{n-1} p$. For example if $P(\text{passing my driving test}) = 0.3$, find $P(\text{passing on the third attempt})$. This will be $0.7^2 \times 0.3 = 0.147$.

What is the probability that I need to take my test at least four times? This can be reworded as: what is the probability of *not* passing on the first or second or third attempt?

$$P(\text{not passing on the first 3 goes}) = 0.3 \times 0.3 \times 0.3 = 0.027$$

Remember the IB use $q=1-p$

and $\binom{n}{x}$ for ${}^n C_x$

The formula book has the following (p 16)

$$\text{If } X \sim \text{Geo}(p) \text{ then } P(X = x) = pq^{x-1}. \quad E(X) = \frac{1}{p} \text{ and } \text{Var}(X) = \frac{q}{p^2}$$

Negative Binomial distribution: This is similar to the geometric distribution, but is defined as the number of trials until the x th success.

For example, how could I find the probability that, when throwing a die, I require eight throws before getting the second 6? The best way to think of this is as follows:

In the first seven throws I must throw one 6.

The eighth throw *must* be a 6.

Use the Binomial distribution for the first part: ${}^7 C_1 \times \frac{1}{6} \times \left(\frac{5}{6}\right)^6 = 0.391$.

$$\therefore P(\text{Second 6 on eighth throw}) = 0.391 \times \frac{1}{6} = 0.065.$$

The formula in the formula book (p 16) is:

$$\text{If } X \sim \text{NB}(r, p) \text{ then } P(X=x) = \binom{x-1}{r-1} p^r q^{x-r} \text{ for } x=r, r+1, \dots. \text{ This models}$$

achieving the ' r 'th success on the ' x 'th go' in a binomial situation.

$$E(X) = \frac{r}{p}, \quad \text{Var}(X) = \frac{rq}{p^2}$$

The cumulative probabilities for a random variable X are given by $P(X \leq x)$. For the Binomial and Poisson distributions these are available on your GDC (see below). But be very careful with the wording of questions, especially differentiating between "at least" and "more than." In these cases we need to take the probabilities of the values we **don't** want away from 1.

Consider how you would work out these probabilities:

- $P(\text{more than 3 are left handed})$
- $P(\text{at least 10 are left handed})$

(Ans. $1 - P(X \leq 3)$ and $1 - P(X \leq 9)$)

1 in 8 people are left handed. What is the probability that, in a group of 12 randomly chosen people, at least 3 are left handed?

"At least 3" means "3 or more." Cumulative probabilities sum from zero upwards. So we require $1 - P(\text{up to 2 left handers}) = 1 - 0.818 = 0.182$.

Using a calculator: On the TI-84 family, Binomial and Poisson distributions will be found in the DISTR menu. "binompdf" will calculate Binomial probabilities, "binomcdf" will calculate Binomial cumulative probabilities. Similarly for Poisson distributions.

Part of a game consists of throwing rubber rings onto a small post. I can usually succeed in 2 throws out of 10. Assuming the same level of success for each throw, calculate the probability that:

- I succeed three times in eight throws.
- I succeed at least three times in eight throws.
- My first success is on my fifth throw.
- My second success is on my fifth throw.

The trick in these questions is to cut through the words and identify the distribution.

- Binomial. $P(3 \text{ in } 8) = 0.147$
- Binomial. $P(\geq 3 \text{ in } 8) = 1 - P(0 \text{ or } 1 \text{ or } 2) = 1 - P(X \leq 2) = 0.203$
- Geometric. $P(\text{1st success on 5th throw}) = 0.8^4 \times 0.2 = 0.08192$
- Negative Binomial.
 $P(\text{2nd success on 5th throw}) = (4 \times 0.2 \times 0.8^3) \times 0.2 = 0.164$

- A bag containing 3 white and 5 black balls has 4 balls withdrawn one at a time in such a way that the first ball is replaced before the next one is drawn. Find the probability of selecting:
 - 3 white balls
 - at most two white balls
 - the first white ball on the fourth draw
 - his third white ball on the fourth draw
- In a game a normal six-sided dice is rolled. What is the expected number of rolls until a 6 is obtained.
- I go to a resort on holiday. At this resort the probability it will rain during the day is 0.2. I resolve to go home at the end of the third day on which it rains.
 - How many days do I expect to stay at the resort?
 - What is the probability I leave at the end of the fifth day?
 - What is the probability I stay at least one week?
- On a particular road in Oxford, 3 in every 10 vehicles is a bus. Assuming this remains true at all times of day, which is more likely – that the eighth vehicle which passes me is the second bus, or that the eleventh vehicle which passes me is the third bus? What further assumption must you make?
- The receptionist at a hotel receives an average of 5.5 enquiries each hour. Assuming the number of enquiries follow a Poisson distribution find the probability that.
 - She receives no enquiries in 1 hour.
 - She receives at least 4 enquiries in 1 hour.
 - She receives fewer than 10 enquiries in 2 hours.

Answers

- 0.132
 - 0.848
 - 0.0916
 - 0.099
- Geo $\left(\frac{1}{6}\right)$, so $E(X) = 6$
- $E(X) = \frac{r}{p} = \frac{3}{0.2} = 15$
 - 0.0307
 - $P(0, 1, 2 \text{ in } 6 \text{ days}) = 0.901$
- 0.0741 and 0.0700. Buses arrive independently of each other.
- 0.0041, 0.798, 0.341.

Note: though I only wrote 3sf for intermediate stages in 3c I calculated the final answer using all the digits from the calculator.

It is very important you do this in the longer questions

Normal Distributions

Two points to note:

- Use cdf, not pdf.
- Normal distributions are defined using variance, but you must enter the standard deviation into the calculator.

Normal distribution:

The Normal distribution is covered comprehensively in the HL revision guide but because of its importance in this option topic we remind you of a few techniques.

The children in a class at school sit an IQ test. The results show that the IQs are Normally distributed with $X \sim N(112, 32.4)$. If a child is chosen at random, find the probability that:

- her IQ lies between 105 and 114.
- her IQ is greater than 120.

Ans. The probabilities are 0.528 and 0.0799.

A normal distribution has a mean of 50 and a variance of 12. Suppose I need to have values only within 1.5 standard deviations of the mean. The boundaries would be $(50 - 1.5\sqrt{12})$ and $(50 + 1.5\sqrt{12})$

If we needed to know how many standard deviations a value (54 say) was from the mean we would reverse the process and using z for the number of standard deviations get $z = \frac{54 - 50}{\sqrt{12}}$.

This generalises to $z = \frac{X - \mu}{\sigma}$.

Invnorm:

In an examination 30% of the candidates fail and 10% achieve distinction. Assuming the marks of the candidates are Normally distributed with mean of 67 and a standard deviation of 10 find the pass mark and the boundary for a distinction.

If we input the probability $X < x$ into Invnorm (x) then it will tell us how many standard deviations below or above the mean x actually is. This is of course always equal to $\frac{X - \mu}{\sigma}$

$\text{Invnorm}(0.3) = -0.524 = \frac{x - 67}{10} \Rightarrow x = 67 - 0.524 \times 10 = 61.8 = 62$ to the nearest whole number.

For the boundary for the distinction (y) we need $P(X < y) = 0.9$

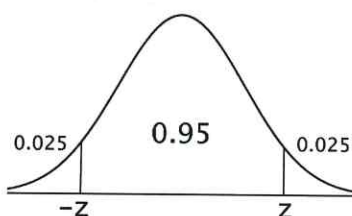
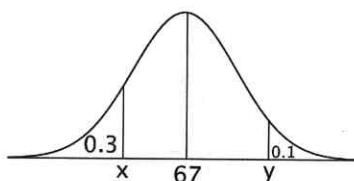
$\text{Invnorm}(0.9) = 1.28 = \frac{x - 67}{10} \Rightarrow x = 67 + 1.28 \times 10 = 79.8 = 80$ to nearest whole number

Note:

- On some calculators you may need to enter $\mu = 0, \sigma = 1$, to get the number of standard deviations (z). These are usually the default settings
- Many calculators will let you put in the mean and standard deviation directly into invnorm to get the final answer without rearranging. The technique above though illustrates aspects of work that come up later in the course.

How many standard deviations from the mean would you need to be to ensure the probability of being outside this range is 0.05 (5%)

From the diagram we can see that z is $\text{invnorm}(0.975) = 1.96$



Cumulative distribution functions

As we have already seen in the case of the binomial and Poisson distribution the cumulative probability distribution for a **discrete random variable**, X , is simply $P(X \leq x)$

I'm not very good at typing. The table below shows the probability distribution X for the number of errors I make in every 100 words

No of errors (x)	0	1	2	3	4
Probability ($X = x$)	0.1	0.24	0.38	0.17	0.11

From this I can construct the cumulative probability distribution which gives $P(X \leq x)$ for each value of x .

No of errors (x)	0	1	2	3	4
$P(X \leq x)$	0.1	0.34	0.72	0.89	1

It is also possible, given a cumulative probability table, to work back to the probability distribution.

For a **continuous variable**, the cdf up to a particular value is the area under the graph from the lower bound up to the value. For example, if the probability density function, pdf, is

$$f(x) = \begin{cases} \frac{12}{7}(x-1)(4-x^2) & \text{for } 1 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

then the proportion of the distribution up to a value x will be given by

$$\int_1^x \frac{12}{7}(x-1)(4-x^2)dx = \frac{24}{7}x^2 - \frac{3}{7}x^4 - \frac{48}{7}x + \frac{4}{7}x^3 + \frac{23}{7}$$

(technically if using x as a limit we should replace the variable within the integral by another letter – it would not cost you any marks if you do as I did and use x for both)

The cdf should then be written out fully as:

$$F(x) = \begin{cases} 0 & \text{for } x < 1 \\ \frac{24}{7}x^2 - \frac{3}{7}x^4 - \frac{48}{7}x + \frac{4}{7}x^3 + \frac{23}{7} & \text{for } 1 \leq x \leq 2 \\ 1 & \text{for } x > 2 \end{cases}$$

As with discrete variables it is possible to work backwards to get the pdf from the cdf

In this case: $f(x) = \frac{dF(x)}{dx}$

Check this works with the function above.

Find $F(x)$ for the random variable, X , which has the following probability density function:

$$f(x) = \begin{cases} 0.4x & 0 \leq x \leq 1 \\ 0.4 & 1 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Hence find the median of X .

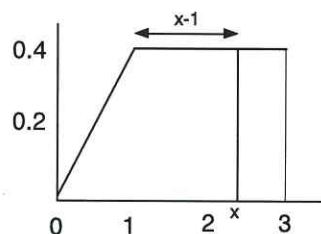
We need to deal with this in two parts. For $0 \leq x \leq 1$,

$$F(x) = P(X \leq x) = \int_0^x 0.4x dx = \left[0.2x^2 \right]_0^x = 0.2x^2$$

For the next region remember that $F(x) = P(X \leq x)$ and this includes the first region whose area is 0.2

$$\text{For } 1 \leq x \leq 3 \quad F(x) = 0.2 + \int_1^x 0.4 dx = 0.2 + \left[0.4x \right]_1^x = 0.2 + 0.4x - 0.4 = 0.4x - 0.2$$

As a check $F(3) = 1$



The median is in the second region so we solve

$$F(x) = 0.5 \Rightarrow 0.4x - 0.2 = 0.5 \Rightarrow x = \frac{7}{4}$$

A continuous cumulative distribution function is defined by

$$F(x) = \frac{x^3}{27} \text{ for } 0 < x < 3.$$

(i) Write down the value of $F(x)$ for $x < 0$ and $x > 3$

(ii) Find the probability density function $f(x)$.

(i) As $F(0)=0$ and $F(3)=1$ it is 0 for all $x < 0$ and 1 for all $x > 3$

(ii) $f(x) = \frac{x^2}{9}$ for $0 < x < 3$ and 0 otherwise

Y is defined as the maximum value of 4 independent random variables all having the distribution above. By first calculating the cumulative distribution function of Y, or otherwise, find the probability density function of Y.

$$F(x) = P(Y < x) = P(\max < x) = P(\text{all four } < x) = \left(\frac{x^3}{27}\right)^4 = \frac{x^{12}}{27^4}$$

$$f(x) = \frac{dF}{dx} = \frac{12x^{11}}{27^4} \quad 0 < x < 3$$

Practice Questions

1. The continuous random variable X has a p.d.f defined by:

$$f(x) = \begin{cases} k(5-x) & \text{for } 0 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

a) Find the value of k

b) Find the cumulative distribution function

c) Find $P(X < 2)$.

2. Find $F(x)$ for the random variable, X , which has the following probability density function:

$$f(x) = \begin{cases} 0.4 & 0 \leq x \leq 1 \\ 0.3 & 1 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

3. If $X \sim \text{Geo}(0.3)$ find the cumulative distribution function for X as a function of x . Hence find $P(X \leq 4)$

Answers

1. a) $\frac{1}{12}$,

b) $F(x) =$

$$\begin{cases} 0 & \text{for } x < 0 \\ \frac{5}{12}x - \frac{1}{24}x^2 & \text{for } 0 < x < 4 \\ 1 & \text{for } x > 4 \end{cases}$$

c) 0.667

2. 0 for $x < 0$
 $0.4x$ for $0 \leq x \leq 1$
 $0.1 + 0.3x$ for $1 \leq x \leq 3$
 1 for $x > 3$

3. $P(X \leq x) = 1 - 0.7^x$
 (prob. of at least one success = 1 - prob. of no successes)
 0.760

The algebra normally deals with variance, but the results are easier to understand in terms of standard deviation.

Expectation Algebra

Linear transformation of a single random variable: Consider the set of values 1, 2, 3, 4, 5. The mean is 3 and the standard deviation is 1.414. If we add 4 to each value, the new mean will be 7, but the standard deviation will remain unchanged (because the spread of results around the mean is exactly the same). However, if we multiply the results by 2, the mean will multiply by 2 and so will the standard deviation; the spread is two times greater than it was. Remember variance = (standard deviation)², so if the standard deviation is multiplied by 2 the variance is multiplied by 2² = 4

Suppose we multiply by 2 then add 4. In summary:

$$X = \{1, 2, 3, 4, 5\}$$

$$E(X) = 3$$

$$\text{Var}(X) = 1.414^2 = 2$$

$$2X + 4 = \{6, 8, 10, 12, 14\}$$

$$E(2X + 4) = 2E(X) + 4 = 10$$

$$\text{Var}(2X + 4) = 2^2 \text{Var}(X) = 8$$

In general, $E(aX + b) = aE(X) + b$ and $\text{Var}(aX + b) = a^2\text{Var}(X)$. Just remember that adding a constant to a random variable has no effect on the variance.

Linear combinations of several random variables:

The general expressions for random variables X and Y with a and b constants are:

$$E(aX \pm bY) = aE(X) \pm bE(Y)$$

Also if X and Y are **independent** then

$$\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) \text{ and } E(XY) = E(X)E(Y).$$

Be careful to distinguish between the random variable $2X$ and the random variable $X_1 + X_2$. For example, when a single die is thrown, let X be the face value. $E(X) = 3.5$, $\text{Var}(X) = 2.917$. If we double the scores, then $E(2X) = 2E(X) = 7$ and $\text{Var}(2X) = 4\text{Var}(X) = 11.67$. However, if you throw the die a second time and add the scores together, each throw is independent, so the total score will be a new random variable, $X_1 + X_2$. $E(X_1 + X_2) = E(X_1) + E(X_2) = 7$, as before, but $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = 5.832$.

An important point to remember is that combinations of Normal distributions are themselves Normal.

An end of year exam consists of a written paper marked out of 80, and an aural test marked out of 10. To calculate the overall result the aural mark is multiplied by 2 and added to the written result. The marks for both the tests are Normally distributed. $W \sim N(68, 7.2)$ and $A \sim N(8.1, 2.2)$ for the written and aural tests respectively. What is the distribution of the overall mark?

In this case, the aural mark has been multiplied by 2, so the distribution is $T = W + 2A$. The mean is calculated as $68 + 2 \times 8.1 = 84.2$, and the variance will be $7.2 + 4 \times 2.2 = 16$. Thus, $T \sim N(84.2, 16)$

Apples are packed into boxes of 24. The weights (in g) of the apples are Normally distributed with $X \sim N(120, 20)$, and the weight of the box (in g) also forms a Normal distribution, $B \sim N(200, 30)$. What is the distribution of the total weight, T , of the box of apples?

$T = (X_1 + X_2 + \dots + X_{24} + B)$. Thus, $E(T) = 24 \times 120 + 200 = 3080\text{g}$, and $\text{Var}(T) = 24 \times 20 + 30 = 510\text{g}$. Thus, $T \sim N(3080, 510)$.

Practice Questions

1. A random variable Y has a mean of 4 and a variance of 1.5. Calculate the expectation and variance of $2Y$, $3Y + 1$, $Y - 3$.
2. Independent random variables S and T are such that $E(S) = 12$, $\text{Var}(S) = 3$, $E(T) = 5$ and $\text{Var}(T) = 1.4$. Find the expectation and variance of $(S + 2T)$ and of $(2S - 3T)$.
3. Over a period of time, the temperatures recorded at a weather station have a mean of 22.4°C and a variance of 2.3°C . Given that the conversion from Centigrade to Fahrenheit is $F = \frac{9}{5}C + 32$ find the expected mean and variance measured in Fahrenheit.
4. If the heights of men (M) are normally distributed $N(175, 100)$ and those of women (W) are $N(150, 81)$ find the probability that a given man is taller than his wife. Assume independence!
Hint: Let $D = M - W$ and find $P(D > 0)$

These are in section 7.2 in the formula books

Note that the variances are added even when the variables are subtracted

Answers

1. 8, 6; 13, 13.5; 1, 1.5.
2. 22, 8.6; 9, 24.6.
3. 72.3, 7.45.
4. $D \sim N(25, 181)$, 0.968
5. $N(1100, 329)$, 0.865
6. $N(15, 134)$, 0.9025

5. A certain type of chocolate bar has a mass which is Normally distributed with mean 50g and standard deviation 4g. A packet contains 20 biscuits. The mass of the packaging is also Normally distributed with mean 100g and standard deviation 3g.
 - a) What is the distribution for the total mass of a packet?
 - b) Calculate the probability that a packet weighs less than 1120g.
6. Assume that the weights in kg of men and women are Normally distributed with distributions $M \sim N(75, 16)$ and $W \sim N(65, 9)$. Calculate the probability that a randomly selected group of six women is heavier than a randomly selected group of five men.

Probability Generating Functions

Section 7.1 of the formula book

The probability generating function (pgf) of a random variable, X , is defined as

$$G(t) = E(t^X) = \sum_{\text{all } x} P(X = x)t^x$$

This has several useful properties

- $G(1) = \sum P(X = x) = 1$
- $G'(t) = \sum P(X = x)xt^{x-1} \Rightarrow G'(1) = \sum xP(X = x) = E(X)$
- $\text{Var}(X) = G''(1) + G'(1) - (G'(1))^2$

Proof

$$G''(t) = \sum P(X = x)x(x-1)t^{x-2} \Rightarrow G''(1) = \sum x^2P(X = x) - \sum xP(X = x) = E(X^2) - E(X)$$

$$\text{Var}(X) = E(X^2) - (E(X))^2 = G''(1) + G'(1) - (G'(1))^2$$

Remember the binomial theorem

$$(a+b)^n = a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \dots + b^n$$

$$= \sum \binom{n}{x} a^x b^{n-x}$$

Find the pgf for a $B(n, p)$ distribution

$$G(t) = \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} t^x = \sum_{x=0}^n \binom{n}{x} (pt)^x (1-p)^{n-x} = (1-p + pt)^n$$

Hence prove that $E(X) = np$

$$G'(t) = np(1-p + pt)^{n-1} \Rightarrow G'(1) = np$$

Find the pgf for a $\text{Geo}(p)$ distribution.

$$G(t) = \sum_{x=1}^{\infty} (1-p)^{x-1} p t^x = \frac{p}{1-p} \sum_{x=1}^{\infty} ((1-p)t)^x = \frac{p}{1-p} \times \frac{(1-p)t}{1-(1-p)t}$$

(using the sum of an infinite geometric series)

$$= \frac{pt}{1-(1-p)t}$$

Some standard PGFs are shown in the table below.

Distribution	$B(n, p)$	$\text{Geo}(p)$	$\text{NB}(r, p)$	$\text{Po}(\lambda)$
PGF	$((1-p) + pt)^n$	$\frac{pt}{1-(1-p)t}$	$\left(\frac{pt}{1-(1-p)t} \right)^r$	$e^{\lambda(t-1)}$

One of the most useful aspects of a pgf is that:

If X and Y are independent then

$$G_{X+Y}(t) = E(t^{X+Y}) = E(t^X)E(t^Y) = G(X)G(Y)$$

The proof follows directly from the fact that if X and Y are independent then $E(X)E(Y) = E(XY)$

If X and Y are independent and $X \sim \text{Po}(n)$ and $Y \sim \text{Po}(m)$ prove that $X + Y \sim \text{Po}(n+m)$

$$G_{X+Y}(t) = G(X)G(Y) = e^{n(t-1)} \times e^{m(t-1)} = e^{(n+m)(t-1)} \text{ which is the PGF for } \text{Po}(n+m)$$

A biased dice has got the probability of p of a six landing face up.

Let X_i be the number of throws it takes to get a six on the i^{th} time I play this game. State the distribution of X_i and $X_1 + X_2 + X_3$

$X_i \sim \text{Geo}(p)$ and $X_1 + X_2 + X_3 \sim \text{NB}(3, p)$ (the total length of time to obtain 3 successes)

Hence write down the PGF for $X_1 + X_2 + X_3$

$$\text{PGF for } X_i = \frac{pt}{1-(1-pt)} \text{ so } X_1 + X_2 + X_3 = \left(\frac{pt}{1-(1-pt)} \right)^3$$

Note: from this we can derive the formula for $\text{NB}(r, p)$ given in the table above.

The probability of a particular outcome occurring is the coefficient of t^r in the expansion of $G(t)$. It can be obtained by expanding the function or by differentiating and setting $t=0$ (illustrated below).

$$G(t) = P(X=0) + P(X=1)t + P(X=2)t^2 + P(X=3)t^3 + \dots \Rightarrow P(X=0) = G(0)$$

$$G'(t) = P(X=1) + 2P(X=2)t + 3P(X=3)t^2 + \dots \Rightarrow P(X=1) = G'(0)$$

$$G^{(2)}(t) = 2P(X=2) + 3 \times 2P(X=3)t + \dots \Rightarrow P(X=2) = \frac{1}{2}G^{(2)}(0)$$

$$\text{and in general } P(X=r) = \frac{1}{r!}G^{(r)}(0)$$

Show that $G(t) = \frac{1}{27}(1+2t)^3$ is a PGF and find the probability $P(X=2)$

by (a) expanding the function (b) by differentiating.

PGF as $G(1)=1$ and each term is positive

$$\text{a) } G(t) = \frac{1}{27}(1+6t+12t^2+8t^3) \text{ hence } P(X=2) = \frac{12}{27} = \frac{4}{9}$$

$$\text{b) } G'(t) = \frac{2}{9}(1+2t)^2 \Rightarrow G''(t) = \frac{8}{9}(1+2t) \quad P(X=2) = \frac{1}{2}G''(0) = \frac{4}{9}$$

Practice Questions

- If X and Y are independent and $X \sim B(n, p)$ and $Y \sim B(m, p)$ prove that $X + Y \sim B(m+n, p)$
- Find the mean and variance of a $\text{Po}(\lambda)$ distribution.
- Find the variance of a $B(n, p)$ distribution
- A PGF for a random variable X is given as $G(t) = \frac{at^2}{(5-4t^2)}$
 - Find the value of a
 - Find $E(X)$
 - Find $P(X=0)$, $P(X=1)$ and $P(X=2)$

Note: In 4b and c it is possible on some calculators to get all these results without finding the derivative explicitly by using the derivative function or equivalent

5. a) Expand $(1 + 2t^2)^5$
- A dice is numbered with 2 ones and 4 threes. Let X represent the score on one roll of the dice.
- b) Write down the probability generating function of X .
I play the game five times. Let Y be my total score on the five games.
- c) Find the probability distribution of Y
6. Two players, A and B are shooting an arrow at a target. The probability A hits the target is $\frac{1}{6}$ and the probability B hits a target is $\frac{1}{4}$. A has the first shot. Let X be the number of shots until the target is hit.
- a) By considering the sums of two infinite geometric series show that the probability generating function of X is:

$$G(t) = \frac{4t + 5t^2}{3(8 - 5t^2)}$$
- b) Hence find the expected number of shots until the target is hit.
7. a) Find the PGF for a random variable X which represents the result of rolling a fair six-sided dice. (In this case trying to simplify the expression by summing it is not very useful – why not?)
- b) Another dice is numbered 2, 4, 6, 8, 10 and 12. What is the PGF for this dice?
- c) In general if X has PGF $G(t)$ what is the PGF for $2X$?

Answers

1. Show pgf is $((1 - p) + pt)^{n+m}$
2. Use $G(t) = e^{\lambda(t-1)}$ and differentiate twice and then use the formula for mean/variance (both = λ)
3. $np(1-p)$
4. a) $G(1) = 1 \Rightarrow a = 1$ b) $E(X) = G'(1) = 10$ c) 0, 0, 0.2
5. a) $1 + 10t^2 + 40t^4 + 80t^6 + 80t^8 + 32t^{10}$ b) $\frac{1}{3}t + \frac{2}{3}t^3$
c) using your answers to part a and b

Y	5	7	9	11	13	15
p	$\frac{1}{243}$	$\frac{10}{243}$	$\frac{40}{243}$	$\frac{80}{243}$	$\frac{80}{243}$	$\frac{32}{243}$

6. b) 4.89 (from GDC)
7. a) $G(t) = \frac{1}{6}t + \frac{1}{6}t^2 + \frac{1}{6}t^3 + \frac{1}{6}t^4 + \frac{1}{6}t^5 + \frac{1}{6}t^6$
b) $G(t) = \frac{1}{6}t^2 + \frac{1}{6}t^4 + \frac{1}{6}t^6 + \frac{1}{6}t^8 + \frac{1}{6}t^{10} + \frac{1}{6}t^{12}$ c) $G(t^2)$

Estimators

Populations often have parameters associated with them that we would like to find out more about (for example the mean and standard deviation of a normal distribution). To do this we collect a sample and measure certain sample statistics and use these to estimate the unknown population values.

We normally use Greek letters for the population value we are trying to estimate and roman letters for the sample estimator.

For example if we were trying to estimate the population parameter, τ , our estimator would be the random variable denoted by T and the particular value of T we obtain from our sample would be t .

Of course there may be many estimators for a single sample so this rule does not apply in every case (we tend to use $\bar{x}$ rather than m as our estimator of μ)

In order to be a good estimator we require firstly that the $E(T) = \tau$. We say that our estimator is **unbiased**.

If we have more than one possible estimator we would choose the one with the smallest variance –so our value of t is less likely to be far away from τ . We say it is a more **efficient** estimator.

If X represents the number of successes in a $B(n,p)$ distribution

show that $\frac{X}{n}$ is an unbiased estimator for p and find its variance.

$$E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p \text{ so unbiased}$$

$$\text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\text{Var}(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Let X be the number of successes in a $B(10,p)$ distribution and Y (independent of X) be the number of successes in a $B(30,p)$ distribution.

Two methods of estimating p are proposed. The first is to pool the two samples and calculate $\frac{X+Y}{40}$, the second is to estimate p from each sample separately and then take the average of the two values

ie $\frac{1}{2}\left(\frac{X}{10} + \frac{Y}{30}\right)$

Show that both these estimators are unbiased and calculate which is the more efficient.

$$E\left(\frac{X+Y}{40}\right) = \frac{1}{40}E(X+Y) = \frac{1}{40}(E(X) + E(Y)) = \frac{1}{40}(10p + 30p) = p$$

$$E\left(\frac{1}{2}\left(\frac{X}{10} + \frac{Y}{30}\right)\right) = \frac{1}{2}\left(\frac{1}{10}E(X) + \frac{1}{30}E(Y)\right) = \frac{1}{2}\left(\frac{1}{10} \times 10p + \frac{1}{30} \times 30p\right) = p$$

$$\text{Var}\left(\frac{X+Y}{40}\right) = \frac{1}{40^2}\text{Var}(X+Y) = \frac{1}{40^2}(\text{Var}(X) + \text{Var}(Y)) = \frac{1}{40^2}(10p(1-p) + 30p(1-p)) = \frac{p(1-p)}{40}$$

$$\text{Var}\left(\frac{1}{2}\left(\frac{X}{10} + \frac{Y}{30}\right)\right) = \frac{1}{2^2}\left(\frac{1}{10^2}\text{Var}(X) + \frac{1}{30^2}\text{Var}(Y)\right) = \frac{1}{2^2}\left(\frac{1}{10^2} \times 10p(1-p) + \frac{1}{30^2} \times 30p(1-p)\right) = \frac{p(1-p)}{30}$$

So the first is the more efficient estimator

Practice Questions

1. X is a random variable with an unknown mean, μ , and variance, σ^2 . Let X_1 , X_2 and X_3 be an independent, random sample taken from X. The following statistics are suggested as estimators for μ :
(1) $\frac{X_1 + X_2 + X_3}{3}$ and (2) $3X_1 - X_2 - X_3$
 - a) Show that both these statistics are unbiased estimators of μ
 - b) Find the variance of each statistic and hence say which is the more efficient estimator.
2. A random sample of two independent variables, X_1 and X_2 are taken from a distribution with mean μ and variance σ^2 . An estimator for μ is given by $aX_1 + bX_2$
 - a) Show that if $aX_1 + bX_2$ is unbiased then $a+b=1$
 - b) Show that $\text{Var}(aX_1 + bX_2) = (2a^2 - 2a + 1)\sigma^2$
 - c) Hence determine the value of a and b for which $aX_1 + bX_2$ has minimum variance.
3. A bag contains beads of which a proportion p are coloured red. A random sample of 5 beads are drawn in turn from the bag, their colour noted and then replaced before the next bead is taken. Let X represent the number of red beads in the sample.
 - a) State the distribution of X
A second sample of n beads is now drawn, with replacement. Let Y be the number of red beads in this sample.
 - b) Prove that both $P_1 = \frac{X}{5}$ and $P_2 = \frac{X - Y}{5 - n}$ are unbiased estimators for p .
 - c) Find in terms of p the variance of both P_1 and P_2
 - d) For which values of n is P_1 a more efficient estimator than P_2

Answers

1. $\text{Var}\left(\frac{X_1 + X_2 + X_3}{3}\right) = \frac{\sigma^2}{3}$, $\text{Var}(3X_1 - X_2 - X_3) = 11\sigma^2$, so first is more efficient
2.
 - a) $E(aX_1 + bX_2) = \mu \Rightarrow a\mu + b\mu = \mu \Rightarrow a + b = 1$
 - b) $\text{Var}(aX_1 + bX_2) = a^2\sigma^2 + b^2\sigma^2$ and replace b with $1-a$
 - c) min of $2a^2 - 2a + 1$ occurs when $a=0.5$ (by differentiating or quadratic curve techniques), hence $b=0.5$
3.
 - a) $B(5, p)$
 - b) $E\left(\frac{X}{5}\right) = \frac{1}{5} \times 5p = p$, $E\left(\frac{X - Y}{5 - n}\right) = \frac{1}{5 - n}(5p - np) = p$
 - c) $\text{Var}(P_1) = \frac{1}{5^2} \times 5p(1-p) = \frac{p(1-p)}{5}$
 $\text{Var}(P_2) = \frac{1}{(5-n)^2} \times (5p(1-p) + np(p-1)) = \frac{(5+n)p(p-1)}{(5-n)^2}$
 - d) solve $\frac{5+n}{(5-n)^2} < \frac{1}{5}$ to get $n > 15$

Sampling

When we wish to find out about a *population parameter* – the mean (μ) in most cases in this course – we need to take a sample from the population and measure the values in the sample and find an estimate ($\bar{x}$) for the parameter. Our *sample statistic* is very unlikely to be exactly the same as the unknown population statistic so in order to work out its significance we need to know the probability distribution of our sample statistic.

Distribution of the sample mean: Using μ and σ^2 for the population mean and variance, and n for the sample size, it can easily be shown by using $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ that $E(\bar{X}) = \mu$ (and hence the estimator is unbiased) and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. It can also be shown (not so easily) that this is in fact the most efficient estimator for μ . Note that as n increases the variance gets smaller so the estimator becomes more efficient (and hence likely to be more accurate).

A sample of size 8 is taken from a population with distribution $N(10,4)$. Find the probability that the sample mean is more than 11.

$\bar{X}$ will be normally distributed (because the parent population is), so $\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{4}{8} = 0.5$. Thus $\bar{X} \sim N(10, 0.5)$.

Then, using your calculator, we find that $P(\bar{X} > 11) = 0.0786$.

Standard error of the mean: The standard deviation of the sample means is $\frac{\sigma}{\sqrt{n}}$. This is sometimes known as *the standard error of the mean*.

Central limit theorem: If the parent population is Normally distributed then so is $\bar{X}$. However, if the parent population is *not* Normal, the sampling distribution of the mean is *still* approximately Normal, and becomes more so the larger the sample size n . This result is called the *Central Limit Theorem*. Generally if n is bigger than 30 we can assume $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ whatever the distribution of X .

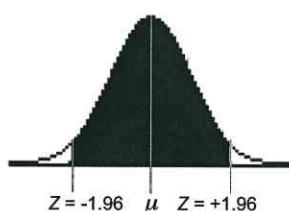
Estimator for σ^2 : We have seen that $\bar{X}$ is an unbiased estimator for the mean, but it would be unusual if knew the value σ^2 at the same time as we were trying to estimate μ . The sample variance is normally denoted S_n^2 , but it can be shown that unfortunately $E(S_n^2) \neq \sigma^2$, so it is not an unbiased estimator. This is because you would expect the *spread* of values in a population to be greater than in a sample. It can be shown that, if S_n is the sample standard deviation, then $\sqrt{\frac{n}{n-1}} S_n$ is an unbiased estimator of the population standard deviation, where n is the sample size. The larger the value of n , the better the estimate.

In summary we use:

$\bar{x}$ as our estimate of population mean (μ) and $s_{n-1} = \sqrt{\frac{n}{n-1}}s_n$ as our estimate for the population standard deviation (σ)

Note the value of s_{n-1} can be obtained directly from the calculator. On a TI-84 s_n is denoted by σ_x and s_{n-1} by s_x

On a Casio they use σ_{on} and σ_{on-1}



Confidence Level	z
90%	1.645
95%	1.96
99%	2.58

The formula book gives the confidence interval as:

$$\bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$$

(section 7.5)

```
ZInterval
Inpt:Data  Stats
σ:2.4
x̄:31.4
n:30
C-Level:.99
Calculate
```

Five measurements of the volume of acid required in a titration are 25.1, 25.2, 25.2, 25.0 and 25.5cm³. Find the mean and standard deviation of the sample and hence obtain estimates for the mean and standard deviation of the volume of acid required.

These results form a sample from which we must estimate the mean and standard deviation of the population of all possible measurements. Check that the mean and s.d. of the five measurements are 25.2 and 0.0280.

Therefore, the required estimates are: $\mu \approx 25.2\text{cm}^3$, $\sigma \approx \sqrt{\frac{5}{4}} \times 0.0280 = 0.0350\text{cm}^3$

Confidence Intervals

Confidence intervals for the population mean with known σ :

A sample mean is an estimator for the population mean, but it is unlikely that the sample mean will be exactly the same as the population mean. We can though give a range either side of a sample mean within which we can be fairly confident the population mean lies: the surer we want to be, the wider the range must be.

The sample means form a Normal distribution with parameters

$E(\bar{X}) = \mu$, $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. To find how many standard deviations

either side of the mean will give us boundaries for 95% of the population we need to calculate $\text{InvNorm}(0.975)$ which comes to 1.96. So we can see that 95% of the sample means lie in the

range $\mu - 1.96 \times \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96 \times \frac{\sigma}{\sqrt{n}}$. Rearranging this gives

$\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}$, and we can be 95% certain

that the population mean lies in this range. Other confidence levels can be used, and their respective Z values are shown on the left.

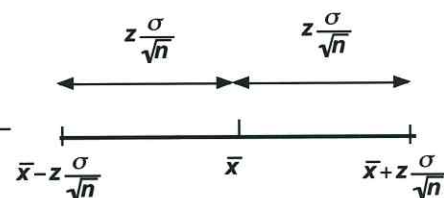
(90% comes from $\text{InvNorm}(0.95)$ and 99% from $\text{InvNorm}(0.995)$)

A plant produces steel sheets whose weights are known to be Normally distributed with a standard deviation of 2.4kg. A random sample of 30 sheets had a mean weight of 31.4kg. Find 99% confidence limits for the population mean.

The limits will be $31.4 \pm 2.58 \times \frac{2.4}{\sqrt{30}} = 31.4 \pm 1.13$. Thus, $30.3 < \mu < 32.5$

The answer can be found directly on most calculators. For the TI-84 family go to **STAT, TESTS, ZInterval**, Choose the **Stats** option and enter the data given in the question. If you are evaluating the confidence interval from data in a list use the **inpt:data** option on the **ZInterval** screen.

The formula above tells us that the width of the confidence interval is $2z \frac{\sigma}{\sqrt{n}}$



A population is known to have a standard deviation of 2cm. How large a sample should I take to be 95% certain that my sample mean is within 0.5cm of the population mean?

We need a maximum of 0.5 either side of $\bar{x}$, therefore $1.96 \frac{2}{\sqrt{n}} < 0.5$ and hence $n > 61.5$ and so the sample must have at least 62 members.

Confidence intervals for the population mean with unknown σ :

In practice, it is likely that the population standard deviation is not known. In such cases we use the same formula as above but replace σ with our best estimate for it, s_{n-1} . Unfortunately this increased uncertainty with regard to our knowledge of the population means that the distribution is no longer normal. Instead it comes from the T-distribution and its actual shape depends on how large our sample is. The parameter we use to define T is the degrees of freedom (ν) which is given by $n-1$. Hence if our sample is of size 10 the distribution of $\bar{X}$ is $T(9)$.

Our confidence interval is now $\bar{x} \pm t \frac{s_{n-1}}{\sqrt{n}}$ (section 7.5 in the formula book)

The value of t is found in a similar way to z when the variance is known. Go to invT and enter 0.975 for a 95% confidence interval. For 9 degrees of freedom this gives 2.26.

For large values of n the t -value will be very close to the z -value.

Tests are being carried out on a new sleeping pill which is given one evening to a random sample of 6 people. The number of hours they sleep may be assumed to be Normally distributed, and is as follows:

7.5 8.1 7.7 6.9 9.2 8.5

Use these data to set up a 95% confidence interval for the mean length of time somebody sleeps after taking one of these pills.

$\bar{x} = 7.983$ hours. The value of $s_{n-1} = 0.806$ (from the calculator). So the statistics we require are:

$$\bar{x} = 7.983$$

$$n = 6$$

$$\nu = 6 - 1 = 5$$

$$s_{n-1} = 0.806$$

From the calculator we find $t = 2.571$, so:

$$7.983 \pm \frac{2.571 \times 0.806}{\sqrt{6}} = 7.983 \pm 0.846$$

So the 95% confidence interval for the mean is 7.14 hours to 8.83 hours.

Again this can be done directly on most calculators (see screen shot for the TI-84 with the data put into list 1)

```
TInterval
Inpt:03:2 Stats
List:L1
Freq:1
C-Level:.95
Calculate
```

```
TInterval
(7.1375,8.8292)
x̄=7.983333333
Sx=.8060190238
n=6
```


Note: Most GDCs want you to input the unbiased estimator, s_{n-1} , (or s_x on the TI-84), so if given the standard deviation you need to convert using

$$s_{n-1} = \sqrt{\frac{n}{n-1}} s_n$$

Practice Questions:

1. A sample of size 6 is taken from a Normal distribution $N(10, 2^2)$. What is the probability that the sample mean exceeds 12?
2. Find the 90% confidence interval for the population mean if a sample of size 20 has a mean of 22.1cm and a standard deviation of 0.8 cm
3. A computer manufacturing company buys large quantities of hard discs from several suppliers. Hard disc quality is checked by a process called RTT which gives results on a continuous scale from 0 to 100. Based on previous experience the company assumes that the results are Normally distributed with a mean of 68 and a standard deviation of 3.
Shipments arrive from suppliers on a daily basis. A sample of 10 hard discs is taken from each shipment at random and tested. If the mean of the sample is more than 67, the shipment is accepted, otherwise it is rejected.
 - a) What is the probability that a hard disc selected at random has a result less than 67?
 - b) Find the probability that a shipment is rejected. (*Hint: Read the question carefully for the conditions of rejection.*)
 - c) There is a \$1000 penalty each day that a shipment is rejected. A particular supplier's hard discs have a mean of 67.5 and a standard deviation of 2.8.
 - i) What is the probability that this supplier's shipment is accepted?
 - ii) What is the expected penalty per 6 day week for this supplier? (*Hint: What is the distribution for the number of days there are rejections in a 6 day week?*).
4. Samples of newly manufactured steel girders are taken to check their lengths. If the standard deviation of the lengths of all girders is 2.6cm, find the smallest sample that must be taken for the standard error of the mean to be less than 0.5cm.
5.
 - a) 50 measurements of the acceleration due to gravity, g , had a mean value of 9.8ms^{-2} and a standard deviation of 0.75ms^{-2} . What are the 95% confidence limits for g .
 - b) How many measurements would be necessary to ensure the length of the 95% confidence interval is less than 0.275ms^{-2} ?
6. Bottles of pickled onions have masses which are Normally distributed with standard deviation 4.6g. The contents of 5 jars taken from a production line are weighed with the following results:
498.2g 501.3g 503.7g 496.8g 502.5g
Calculate a 95% confidence interval for the mean.
7. Families in the state of Ibonia generally have large numbers of children. The government carries out a survey of 100 families and asks each of them how many children there are in the family. The results were:

No of children	0	1	2	3	4	5	6
No of families	8	22	27	19	13	8	3

E

- a) Calculate estimates of the mean and variance of the number of children per family in the population as a whole.
- b) Calculate a 90% confidence interval for the mean number of children per family. (*Hint: If using your GDC, enter L_2 for Frequency.*)

Answers

1. 0.0072
2. (21.8, 22.4)
3. 0.369, 0.146, 0.714, \$1716
4. 28
5. [9.6, 10.0], 217
6. [496.5, 504.5]
7. 2.43, 2.288, [2.18, 2.68]

Significance Testing

Testing a mean: Suppose we believe the population mean (μ) is a particular value and wish to test this belief (our *null hypothesis*). We would collect our sample, measure the sample mean and see how close it is to our supposed value. We would have to reject our null hypothesis if our sample mean was a long way from this value. We measure 'a long way' by how many standard deviations it is from μ . We would normally reject our null hypothesis if the chances of it being that far away or further is less than 5% if the mean really was what we think it is. For a *two-tailed test* this corresponds to 1.96 standard deviations (if population variance known)

Typical examples:

- Patients have a 60% chance of recovering from a particular illness. A new drug is tested on 100 patients and 68 recover. Could this have happened by chance? If we find it is very unlikely then we can accept the fact that the drug is having a positive effect.
- A cereal manufacturer claims that his packets contain a mean weight of cereal of 200g. A rival samples 50 packets and finds the mean weight is only 190g. If this is significantly less than the nominal mean, then this will cast doubt on the manufacturer's claim.

Known σ : The following example takes you through the process stage by stage:

The IQ scores of a population are Normally distributed with a mean of 100 and a standard deviation of 15. A psychologist wishes to test the theory that eating chocolate improves your score. A random sample of 60 people are selected and they are each given a bar of chocolate before sitting an IQ test. Their mean score is 103. Test the psychologist's theory at the 5% level.

Stage 1: Set up the hypotheses.

The *null* hypothesis is the status quo; in this case, that the mean IQ of the 60 people is 100. The *alternative* hypothesis is that the chocolate is having an effect, and that the mean score has increased. Write the hypotheses like this:

$$H_0: \mu = 100$$

$$H_1: \mu > 100$$

Stage 2: Calculate the probability.

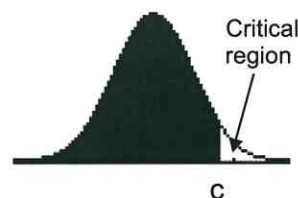
The diagram on the right shows the "unlikely" region (ie the top 5%) of a Normal curve. The value where this occurs is called the *critical value*, and the region is called the *critical region*. We could find this critical value and see whether 103 was in the critical region or not, (We need to consider this later) but for now we will do the calculation directly on the GDC.

Among the results obtained is the p-value, $p=0.0607$. This is the likelihood of getting this sample mean if H_0 is really true.

As 6.1% is greater than 5% the result is not significant.

Stage 3: State the conclusion.

103 is *not* in the critical region (a mean that high could have happened by chance), so we accept H_0 and conclude that there is no strong evidence to support the psychologist's claim.



On a TI-84 you can find Z-Test under the STATS, TESTS menu.
In this case you want $> \mu_0$

One-tailed or two-tailed? In the example above, the claim was that chocolate increased IQ, so we were only testing if the mean was higher than normal. This was a *one-tailed test*. However, if the claim was that chocolate affected IQ (either way), then H_1 would have been $\mu \neq 100$, and the test would be *two-tailed*. In practice, the only difference is that the 5% critical region is split into 2.5% at each end. If doing the test on a GDC this is taken care of by the calculator.

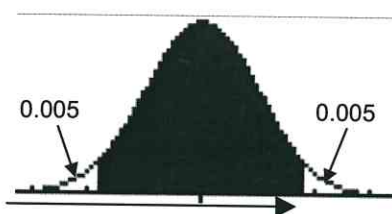
A machine produces bolts with a mean diameter which is normally distributed such that $X \sim N(0.580, 0.015^2)$. After servicing, an operator suspects that the mean diameter has changed. A sample of 50 bolts is found to have a mean diameter of 0.574cm. Assuming the standard deviation is unchanged test, at the 1% level, whether this sample provides evidence that the mean diameter *has* changed.

We are not asked to test whether the diameter has decreased or increased, only whether it has changed – so we will carry out a 2-tailed test.

$$H_0: \mu = 0.580$$

$$H_1: \mu \neq 0.580$$

The GDC gives a p-value of 0.00468, and since this is less than 0.01 we must reject H_0 . There is evidence to suggest that the machine is not producing bolts with mean diameter 0.580cm.



The critical values can also be found using the confidence interval tools on a gdc but you must enter the value of μ assumed under H_0 as the mean, and not $\bar{x}$

Finding the critical values: An alternative way of tackling the question above, and one you will have to do on occasions, is to find the *critical* values; these are the boundary values where the "likely" region becomes the "unlikely" region. The unshaded areas under the Normal curve on the left represent the critical regions for a 2-tailed test with significance level 1%. The arrow underneath shows that we require the z value for an area of 0.995. The calculator gives $\text{InvNorm}(0.995) = z = 2.576$. As always z is the number of standard deviations from the mean so the critical values

$$\text{are } \mu \pm 2.576 \frac{\sigma}{\sqrt{n}} = 0.580 \pm 2.576 \times \frac{0.015}{\sqrt{50}} = 0.580 \pm 0.00546 \text{ resulting}$$

in critical values of 0.57454 and 0.58546. The sample mean of 0.574 is therefore just inside the left hand critical region, hence the

significant result. Compare this formula $\mu \pm z \frac{\sigma}{\sqrt{n}}$ with the one for

the confidence intervals given in the formula book $\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$

Note the gdc will also give the number of standard deviations (z) the sample mean is from the assumed population mean, in this case -2.83

Unknown σ : At the bottom of page 15 we saw how to calculate confidence intervals when the population standard deviation is unknown by using the *t*-distribution. We carry out a significance test in pretty much the same way.

A data checker claims he can process database records at an average rate of 1 every 10 minutes. He is timed checking 5 records, and manages the following results (in minutes):

9.88 10.18 10.23 10.39 10.25

Assuming the times are Normally distributed, test at the 5% level whether his claim is likely.

$$H_0: \mu = 10$$

$$H_1: \mu \neq 10$$

Sig. Lev. 5%

From the gdc we get $t=2.21$ (so our mean is 2.21 s.d.s from the mean) and $p=0.0913$.

The p-value is greater than 0.05 so the result is not **significant** and we therefore accept H_0 . In other words, there is no evidence to suggest that the mean time is *not* 10 minutes.

The gdc screen will also give you $\bar{x}$ and s_{n-1} . It is worth writing these down also – even though they are not needed for the test because you never quite know what the marks are for in the exam and they might even provide some follow through marks if you make an earlier error.

The formula for the critical values is $\mu \pm t \frac{s_{n-1}}{\sqrt{n}}$ (again compare with

the confidence interval formula in section 7.5 of the formula book). The t-value is found once again by $\text{invT}(0.975)$ with $5-1=4$ degrees of freedom. Once again, we could avoid all unnecessary calculations and obtain the critical values by using the confidence intervals tool of the gdc. Enter the data from the previous question, then use the T-Interval test as before (use STATS and change $\bar{x}$ to 10. Put in a confidence level of 0.95 (you would have put in 0.90 if it were a 1-tailed test to get a 'tail' of 0.05, and just ignore the second value). The critical values turn out to be 9.77 and 10.23. $\bar{x} = 10.19$ so is within these limits, and the result is therefore not significant.

Test for Differences Between Means. Paired Sample:

Suppose we have two sets of data $X_1, X_2 \dots X_n$ and $Y_1, Y_2 \dots Y_n$ which are paired in some obvious way and we wish to test

H_0 : The two means are equal against H_1 : The two means are not equal

We first of all find the differences in the two values for each member of the population and then test that the average difference for the population (μ_D) is equal to zero. σ^2 is hardly ever known so the differences D have a T distribution with $n-1$ degrees of freedom. $\bar{d}$ is the mean of the differences and s_{n-1} the unbiased estimator of standard deviation of the difference. Our hypotheses

become $H_0: \mu_D = 0$ and $H_1: \mu_D \neq 0$ Our test statistic is $t_{\text{test}} = \frac{\bar{d}}{\frac{s_{n-1}}{\sqrt{n}}}$

Five candidates attended a special training course to improve their job skills. They were tested before the course started and were tested again at the end of the course. The results were as follows.

Candidate	1	2	3	4	5
Score before course	107	93	105	96	92
Score after course	115	94	107	94	98

Assuming that the test does measure the necessary skills taught in the course, determine whether the course improved the candidates test performance at the 5% level.

$H_0: \mu_D = 0$ and $H_1: \mu_D > 0$ The differences are 8 1 2 -2 6

putting these into a gdc gives us $t=1.68$, $p=0.0844$, $\bar{d}=3$, $s_{n-1}=4$

From the p-value we can see that the result is not significant so there is no reason to reject the null hypothesis that there has been no improvement.

Type I and Type II errors: We can never be *certain* of a result when carrying out a significance test. It is possible that a significant result leads us to reject H_0 when H_0 is in fact true. It is also possible that a significant result leads us to accept H_0 when, in fact, H_1 is true. The first of these errors is called *Type I* and the second is called *Type II*.

Type I errors are easy to deal with. If we choose a 5% significance level, this means that a significant result will happen *by chance alone* 5% of the time. Thus there is a 5% probability of a Type I error occurring. In general, the probability of a Type I error is the same as the significance level. So why don't we make the significance level really small? Because the less chance there is of a Type I error, the greater the chance of a Type II error.

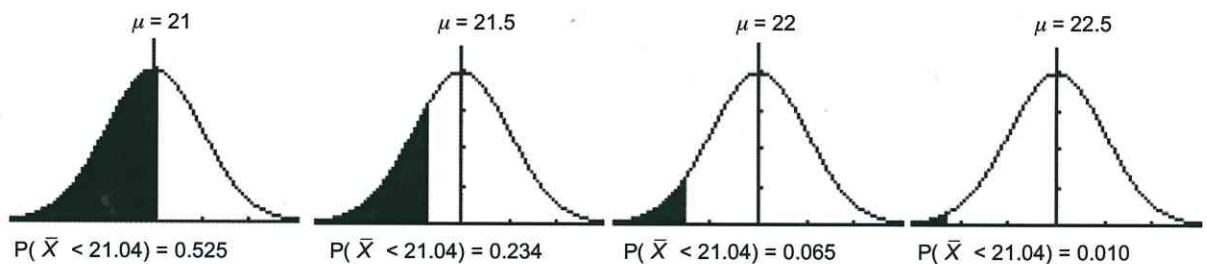
Now, a Type II error is rather more tricky. We want to know the probability that a result is *not* in the critical region if the population mean is actually different from what we think it is. But without knowing that mean, we can't do the calculation. And, if we knew the mean, there wouldn't be any point doing the significance test in the first place! The diagrams below may help to explain this contorted piece of reasoning.

Suppose we carry out a 5% significance test with $H_0: \mu = 20$ and $H_1: \mu > 20$. $\sigma = 2$ and $n = 10$. The critical value is $20 + 1.65 \times \frac{2}{\sqrt{10}}$ and hence the critical region is $\bar{x} > 21.04$.

Suppose the mean is not actually 20. If we get less than 21.04 we will mistakenly accept the hypothesis that it is and commit a type II error. How likely is it that we get $\bar{x} \leq 21.04$ for different values of μ ?

The shaded areas below each show the probability that we will *not* get a significant result from our test – the Type II error. You can see how they vary. So without knowing an alternative for μ there is little point in calculating a Type II error.

In practice, we would carry out more than one test to increase our chances of reaching the right conclusion.



The procedure for finding a Type II error consists of two stages:

- 1 Find the critical region under the assumption H_0 is true
- 2 Find the probability of **not** being in the critical region using the alternative value of the mean (or other parameter being tested)

Boxes of breakfast cereal have contents whose masses should form a Normal distribution $M \sim N(375, 15^2)$. It is thought that the machine which fills the boxes has a defect, and the mean is no longer 375g. The contents of a random sample of 16 boxes are weighed and a significance test at the 5% level carried out.

- a) Write down null and alternative hypotheses.
 - b) Given that the actual value of μ is 380g, find the probability of making a Type II error.
 - c) How could the test be changed so as to reduce the probability of a Type II error?
- a) $H_0: \mu = 375$
 $H_1: \mu \neq 375$
- b) First find the critical regions.
 This is a two-tailed test, so the Z value can be found from the calculator using $\text{invNorm}(0.975) = 1.96$. Thus, the critical values are $375 \pm 1.96 \frac{15}{\sqrt{16}} = (367.65, 382.35)$. So the alternative hypothesis will be accepted if $\bar{X} < 367.65$ or $\bar{X} > 382.35$.
 Now find the probability of **not** being in these regions if $\mu = 380$
 With $\bar{X} \sim N\left(380, \frac{15^2}{16}\right)$, $P(367.65 < \bar{X} < 382.35) = 0.734$, which is therefore the probability of a type II error.
- c) Increasing the significance level to 10% will increase the size of the critical region, and hence the probability of getting a sample mean within that region. A larger sample size would also reduce the probability of a Type II error.

In fact, the probability is lowered to 0.621. Try it.

We can find Type I and Type II errors for any distributions. The method follows the same two rules as above.

It is thought that in a particular university the probability a person owns a car is $p = 0.3$. It is decided to take a sample of 12 people to test this belief and to reject $p = 0.3$ in favour of p being greater than 0.3 if, in the sample, more people than expected were found to own a car.

- a) State the distribution of the number of people in the sample who own a car (X) assuming $p = 0.3$
 - b) Find the critical region for rejecting at the 5% significance level the hypothesis that $p = 0.3$ against the belief that it is in fact larger than this.
 - c) For the above test calculate the probability of a Type II error when the population proportion is in fact (i) 0.4 (ii) 0.5
- a) $X \sim B(12, 0.3)$
- b) $H_0: p = 0.3$ $H_1: p > 0.3$ We firstly need the critical region. We reject H_0 if we get too many people in our sample who own a car, ie if $X \geq c$. At the 5% significance level we need $P(X \geq c) < 0.05$ Using the binomial function on the gdc we get $P(x \leq 5) = 0.882$, $P(x \leq 6) = 0.961$ Hence reject H_0 if $X > 6$ ie $X \geq 7$
- c) Probability of a Type II error is the probability of being in the critical region if $p = 0.4$
 $P(X \geq 7 | p = 0.4) = 1 - P(X \leq 6 | p = 0.4) = 0.158$

Practice Questions

Answers

1. $P(\bar{x} > 1012) = 0.043$, so (a) is no and (b) is yes.
2. $t = -4.57$ which is significant. Pineapples seem to be smaller. Assume sample is not all from same batch.
3. $H_0: \mu_D = 0$, $H_1: \mu_D > 0$
 $p = 0.230 > 0.05$ so no reason to reject H_0
4. a) Reject if

$$\bar{x} > 165 + 1.645 \frac{5}{\sqrt{10}} = 167.6$$
 b) $P(\bar{X} < 167.6 \mid \mu = 1.75) = 0.0694$
5. a) $P(A \text{ wins all races} \mid p = 0.5) = 0.5^3 = 0.125$
 b) $P(B \text{ wins at least one} \mid p = 11/15) = 1 - P(A \text{ wins all}) = 1 - \left(\frac{11}{15}\right)^3 = 0.606$

1. A machine is designed to produce rope with a breaking strain of 1000N. The standard deviation is known to be 21N. A random sample of nine pieces of rope had a mean breaking strain of 1012N.
 - a) Is there evidence at the 5% significance level that the breaking strain has changed?
 - b) Is there evidence at the 5% significance level that the breaking strain has *increased*?
2. A fruit buyer believes that the pineapples he buys from a particular grower are not as big as they used to be. He claims that the pineapples used to average 0.75kg each. A random sample of 20 pineapples gives the following weights in kg:
 0.65 0.68 0.77 0.71 0.67 0.70 0.70 0.72 0.76 0.73
 0.75 0.69 0.72 0.73 0.69 0.71 0.70 0.69 0.75 0.78
 Carry out a test at the 5% significance level and state the conclusion, including any assumptions you have made.
3. It is decided to test whether or not a lack of sleep reduces a person's reaction time. To do this the reaction times of 8 people are measured in the morning and again eighteen hours later during which time the person has not slept. These times are recorded in the table below. Use this data to test at the 5% level whether 18 hours without sleep reduces a person's reaction time.

Person	A	B	C	D	E	F	G	H
First test	0.5	0.6	0.5	0.8	1.2	0.6	0.3	0.5
Second test	0.6	1.2	0.4	0.5	1.1	0.8	0.5	0.5
4. I thought I had read somewhere that the average height of students in year 11 is 1.65cm but my friend says I am not remembering correctly and it is in fact more like 1.75cm. We decide to measure a randomly selected group of 10 students in year 11 in our school. We agree to take the standard deviation to be 5cm and to use my prediction as the null hypothesis.
 - a) Find the critical region for the test at a 5% significance level.
 - b) Find the probability of a Type II error if my friend's prediction is in fact correct.
5. Two track athletes A and B have met 15 times in the same event. A had the faster time on 11 occasions, but now B claims there is nothing between them. At a forthcoming meeting they will be competing together in 3 races and it is decided to test whether B is the same standard as A (If p = probability A wins, $H_0: p = 0.5$) or whether the previous results are more valid as a measure of their ability ($H_1: p = 11/15$). H_0 will be accepted if B obtains a faster time than A in any of the three races otherwise H_1 will be accepted. Find the probability of making a) a Type I and b) a Type II error.

Bivariate Data

As its name suggests bivariate data links two paired quantities and is often shown by way of a scatter diagram, eg heights and weights of people in a population.

The Product Moment Correlation Coefficient

If we plot our data on a scatter diagram we can measure how closely the data is to a straight line by finding the correlation coefficient. If all the data is perfectly along a line with a positive gradient it will have a correlation coefficient of 1. If the line has negative gradient it will be -1. If there is no correlation at all (ie X and Y are independent) the value will be 0.

If we have a population and some bivariate data associated with it then there will be a population correlation coefficient (ρ). To estimate ρ we might take a sample and work out the correlation coefficient for the sample (r). Like all estimators this will be a random variable represented by R and we would hope that $E(R) = \rho$

For bivariate data following the two distributions X and Y , ρ is defined as

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Where $\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$ and μ_X, μ_Y are the population means of X and Y respectively – therefore $E(X) = \mu_X$ and $E(Y) = \mu_Y$. You need to be able to prove that ρ satisfies the two conditions stated above.

Prove that if X and Y are independent then $\rho=0$

First show that $E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X\mu_Y$

$$E[(X - \mu_X)(Y - \mu_Y)] = E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) = E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y$$

$$= E(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y = E(XY) - \mu_X \mu_Y$$

Hence $\text{Cov}(X, Y) = E(XY) - \mu_X \mu_Y = E(X)E(Y) - \mu_X \mu_Y$ (if X, Y are independent) $= \mu_X \mu_Y - \mu_X \mu_Y = 0$

Prove that if X and Y are linear related ie $Y=aX+B$ then $\rho=1$

Using $\text{Cov}(X, Y) = E(XY) - \mu_X \mu_Y$ we obtain

$$\begin{aligned}\text{Cov}(X, Y) &= E(X(aX + b)) - E(X)E(aX + b) = E(aX^2 + bX) - E(X)E(aX + b) \\ &= aE(X^2) + bE(X) - aE(X)^2 - bE(X) = aE(X^2) - a(E(X))^2 = a\text{Var}(X)\end{aligned}$$

$$\text{Also } \text{Var}(Y) = \text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$\text{Hence } \rho = \frac{a\text{Var}(X)}{\sqrt{a^2(\text{Var}(X))^2}} = \frac{a}{\sqrt{a^2}} \text{ as the denominator must be positive we}$$

get $\rho = 1$ if $a > 0$ and $\rho = -1$ if $a < 0$

The Sample Correlation Coefficient:

The sample correlation coefficient (R) is defined as:

These equations are in section 7.7 of the formula

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)\left(\sum_{i=1}^n (Y_i - \bar{Y})^2\right)}} = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2\right)}}$$

It is given that this is an unbiased estimator of ρ

Particular values of r can be worked out from given data using this formula (with small x 's, y 's replacing the capital letters). Fortunately you are unlikely to be asked to use these in an exam as the calculator will work the value out for you.

On a TI-84 make sure you have diagnostics ON (found under the CATALOG menu (above 0))

Enter the data and go to STAT TEST LinRegTTest which provides all the data needed for all of this section of the course

For the data below work out the sample product moment coefficient, by using the formula by using the inbuilt function on the calculator.

Student	A	B	C	D
Score in Maths	55	85	42	67
Score in Science	62	82	58	65

Comment on your answer

We put the values into 2 lists and obtain $r=0.950$. There therefore appears to be a very strong positive correlation between the two variables.

Testing the hypothesis that $\rho=0$

If X and Y can be assumed to come from normal distributions it is possible to assess whether or not the two variables can be considered to be independent.

If you are given the actual values in your sample some calculators will do the test directly.

Following the same process as above to go to LinRegTTest and set $\rho \neq 0$. This will give you the p -value and, as usual, you would normally reject H_0 if this is less than 0.05.

Using the data above the p -value = 0.0503, so the result is not significant at the 5% level and there is not enough evidence to reject H_0 that the two variables are independent. (though the correlation is strong there is not enough data to be sure – so more values need to be taken)

In exam questions where just the r value is given and not the original data we have to use the following fact

$R\sqrt{\frac{n-2}{1-R^2}}$ has the t -distribution with $(n-2)$ degrees of freedom so for a particular value of r our test statistic (the number of standard deviations r is from 0) is given by $t = r\sqrt{\frac{n-2}{1-r^2}}$

Suppose a sample of size 10 is taken and $r=0.14$. Test at the 5% level that

$H_0: \rho=0$ against $H_1: \rho \neq 0$

Our test-statistic is therefore $0.14 \sqrt{\frac{8}{1-0.14^2}} = 0.400$

This is a measure of how many standard deviations our value is from the mean.

On a two-tailed test the critical points for this can be found from $\text{InvT}(0.975)$ using 8 degrees of freedom and are equal to ± 2.31

From this we see our result is not significant so there is no reason to reject the null hypothesis that $\rho = 0$

Twenty runners compete in two races, the first over 100m and the second over 1500m and the times recorded. The product moment correlation for the sample was found to be 0.25.

Assuming that the data can be considered as a random sample from a bivariate normal distribution with correlation coefficient ρ , test at the 5% level the hypothesis that there is a positive correlation between the times taken by runners generally over the two distances.

$H_0: \rho = 0$, $H_1: \rho > 0$

$$t = r \sqrt{\frac{n-2}{1-r^2}} = 0.25 \sqrt{\frac{18}{1-0.25^2}} = 1.0954$$

Critical value for a one-tailed test on 18 degrees of freedom is $\text{InvT}(0.95) = 1.734$

The result is not significant so there is no reason to reject the null hypothesis that $\rho = 0$

Linear Regression

If the correlation coefficient suggests there is a reasonably good linear relationship between the two variables it is worth trying to find the line of best fit. The usual line chosen is the line that minimises the squares of the distances from the line to the points. This is called the least squares regression line and can be regarded as the 'line of best fit'.

If it minimises the vertical distance we get the line Y on X and if we minimise the horizontal distances we get the line X on Y .

It is important to note that the line of Y on X should only be used for predicting values of Y for given values of X within the range of the data (interpolation) and not for values outside this range (extrapolation).

If you want to predict values of X from a value of Y , you must use the line X on Y .

There are formulae for both these lines in the formula books (section 7.7) but you would normally be expected to simply use the calculator.

Find the line y on x and x on y for the following data:

Student	A	B	C	D
Score in Maths (x)	55	85	42	67
Score in Science (y)	62	82	58	65

On the TI-84 LinRegTTest menu you are given the option to put your regression equation in as one of your y equations. If you need to calculate values it is a good idea to enter Y_1 here.

The last part can then be done by typing $Y_1(60)$

The procedure is the same as for the correlation coefficient. Put the data into lists and go to the linear regression menu

This gives the line y on x as $y=0.549x+32.6$

To get x on y reverse the 2 lists to obtain $x=1.64y-47.5$

Find the value of y when x is 60

$$y=0.549x60+32.6 \approx 65.5$$

Practice Questions

1. For the data below

x	25	30	35	40	45	50
y	73	65	66	53	43	43

- Find the correlation coefficient
- Test the hypothesis that the two factors are independent (ie $\rho = 0$)
- Calculate the equation for the regression line y on x
- Calculate the equation of the regression line x on y
- Estimate the value of y when $x=37$
- Estimate a value of x when $y=49$
- Explain why it is not possible to use your equations to estimate a value of y when $x=65$

2. It is suspected that there is a correlation between the amount of sunlight absorbed by plants in an experiment and the height to which they grow.

A test on 25 plants yielded a correlation coefficient of 0.32. Test at the 5% level whether there is evidence that there is a positive correlation between the amount of sunlight and the height of a plant.

3. It is felt that there might be a correlation between the distance a student lives from school and the number of after school clubs he/she attend. A sample of 20 people was taken and the correlation coefficient was found to be -0.12. Test at the 5% level the hypothesis that there is no correlation against the alternative hypothesis a) there is a correlation and b) there is a negative correlation.

Answers

- 0.962
 - $p=0.00209 < 0.05$ so reject H_0
 - $y=106-1.31x$
 - $x=78.0-0.708y$
 - 57.8
 - 43.2
 - $x=65$ is outside the range of the data given.

$$2. t = 0.32 \sqrt{\frac{23}{1-0.32^2}} = 1.62$$

critical value for $T(23) = 1.71$
Result is not significant so no reason to reject H_0

- $t = -0.957$
 - $\text{inv}T(0.975) = 2.10$
 - $\text{inv}T(0.05) = -1.73$

Result is not significant so no reason to reject H_0